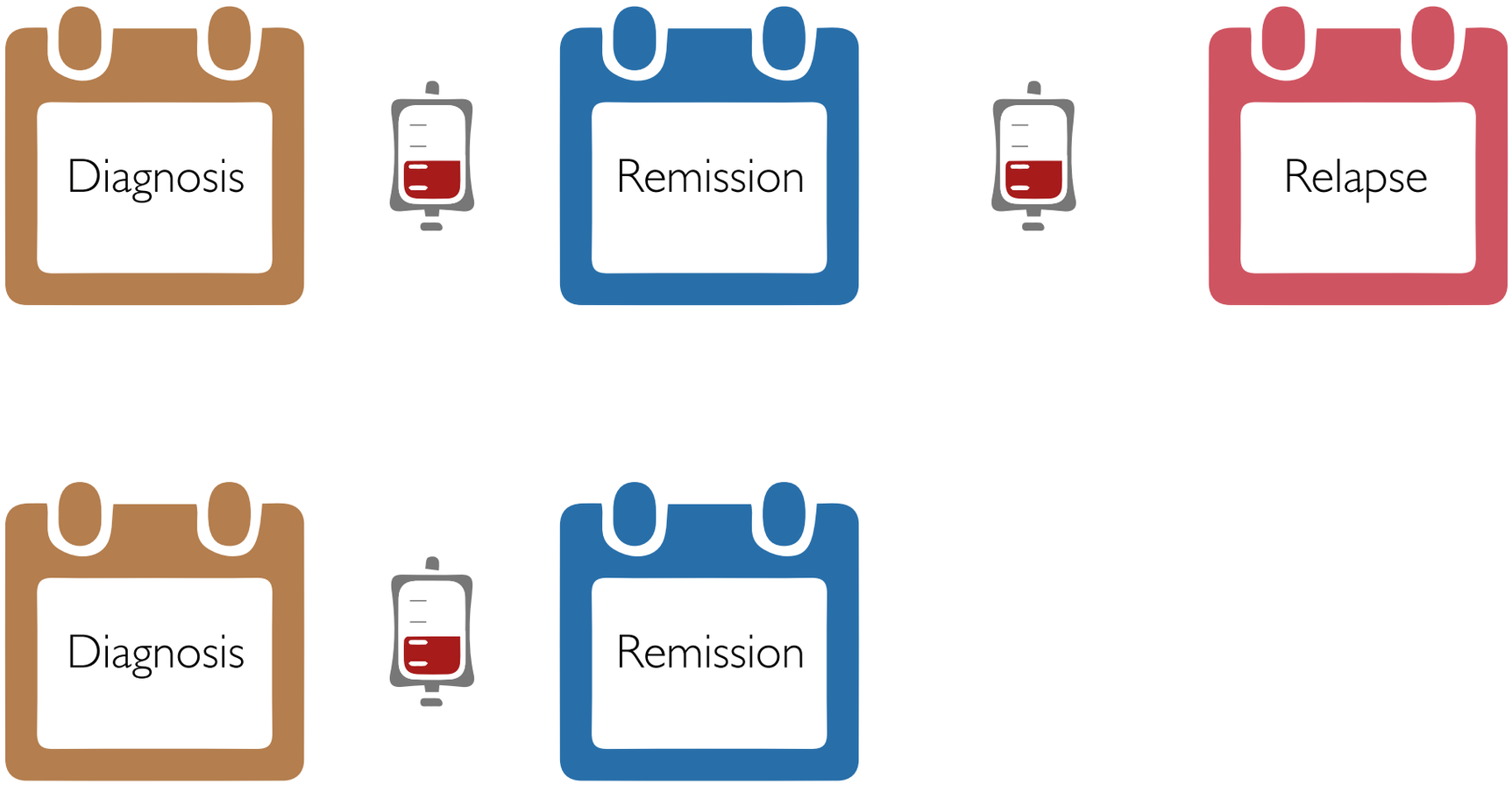
Get your variants together!
Call and combine mutations over time from Leukemia RNA-Seq

Anna Quaglieri

Bioinformatics Seminar 21st May 2019

Data available:
Leukemia RNA

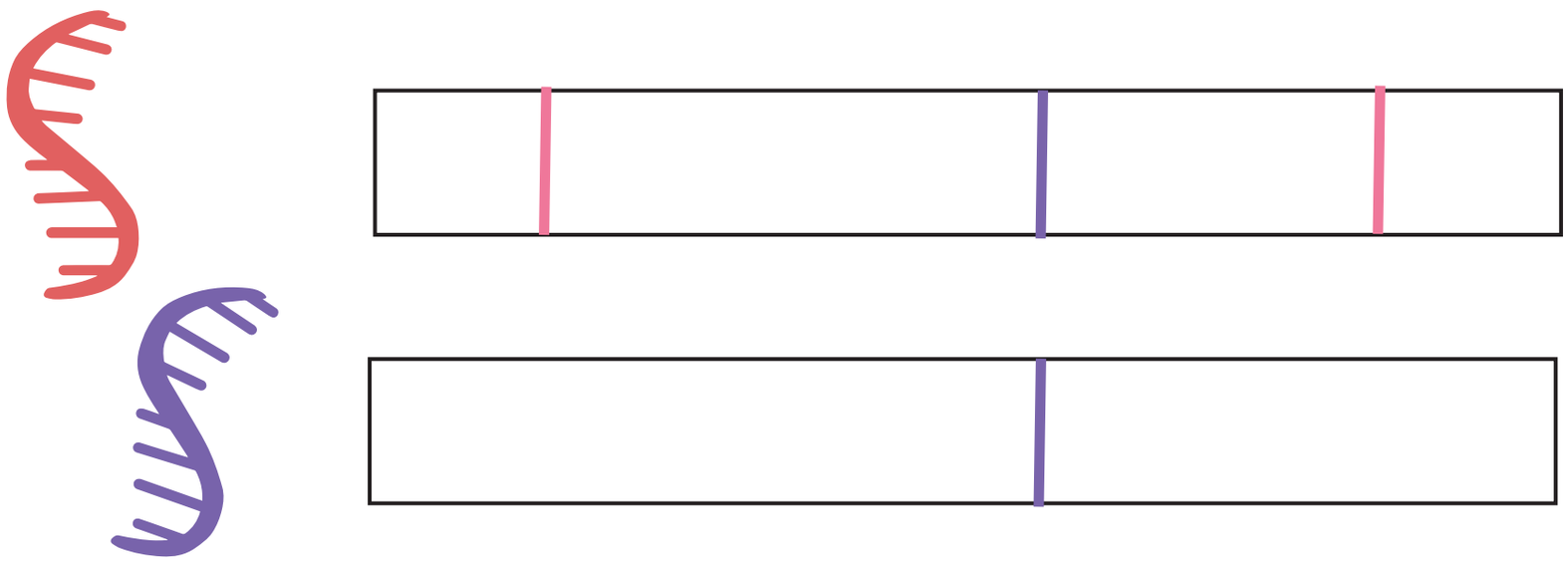
Data available:
Leukemia RNA



We needed to extract all things from RNA!

Calling **tumour-only** variants in **RNA-Seq** with samples **over time**

Tumour-normal variant calling



Tumour sequence

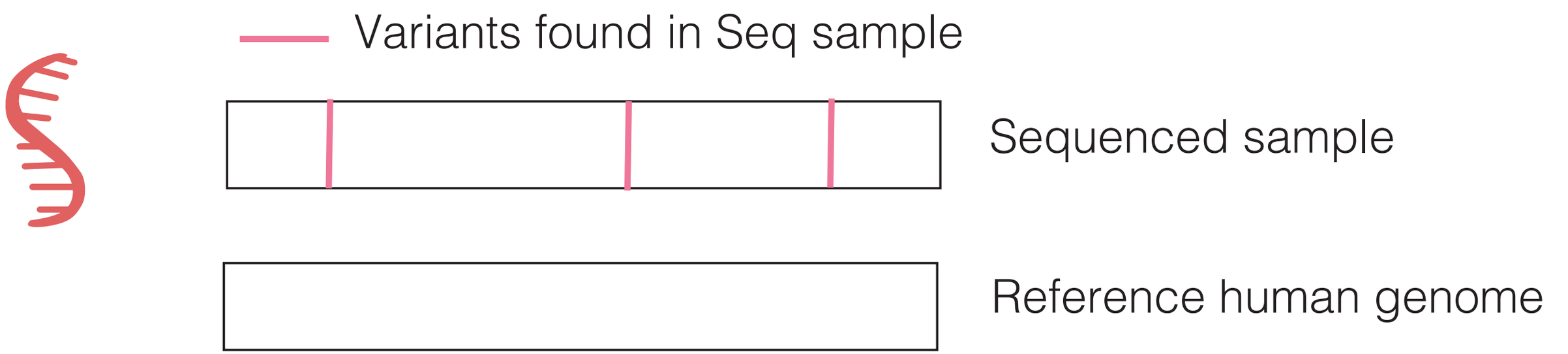
Normal sequence

Reference human genome

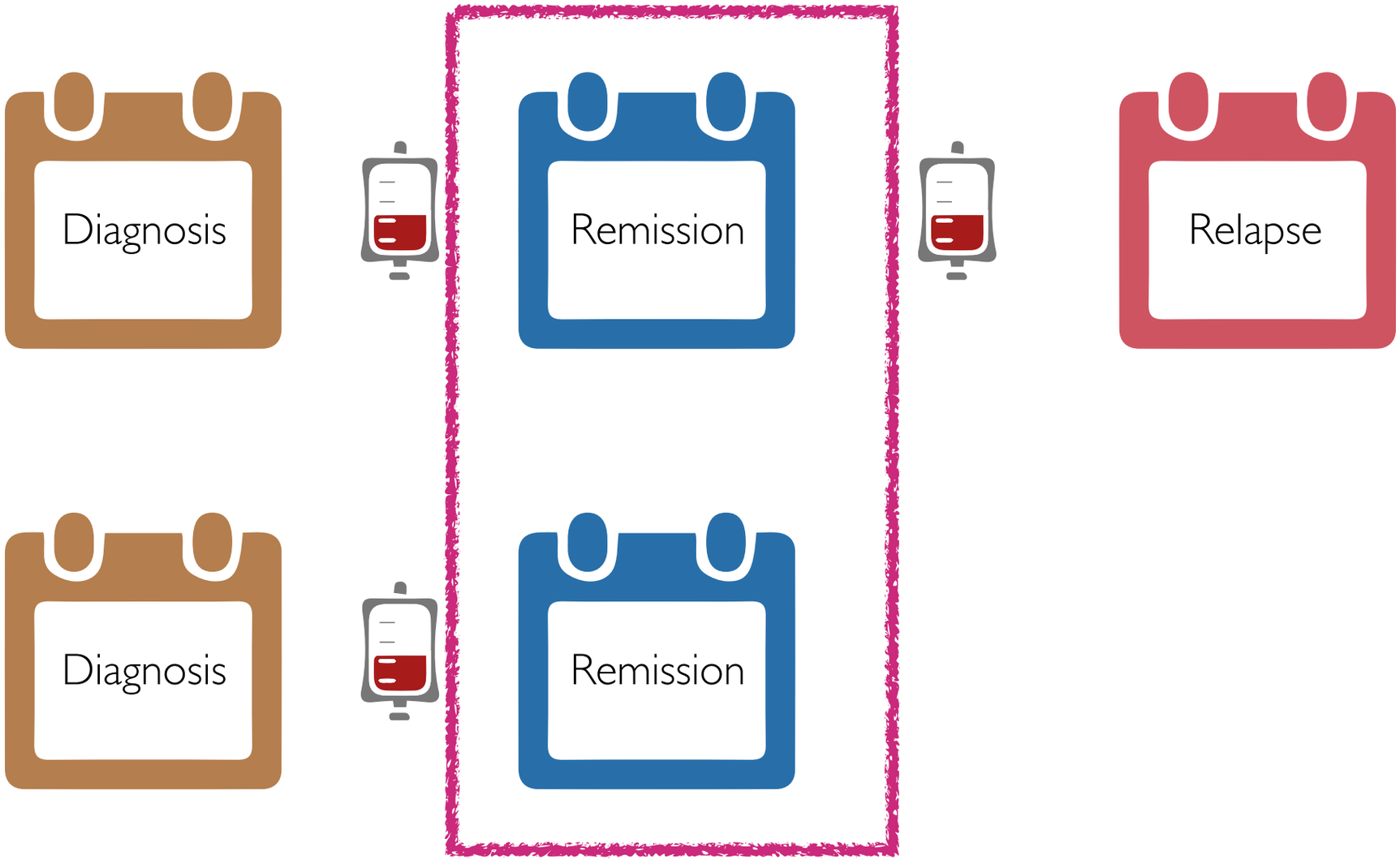
Somatic variants found only in the tumour sample

Germline variant found in both the tumour and normal samples

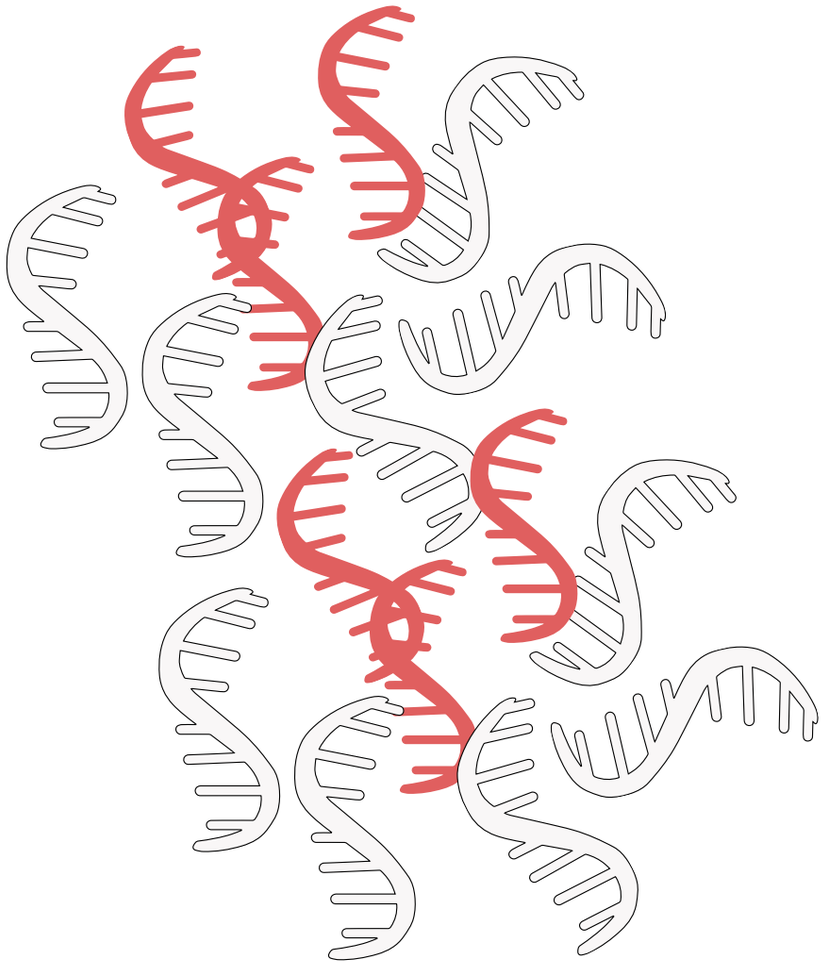
Tumour-only variant calling



Sometimes not reliable normals

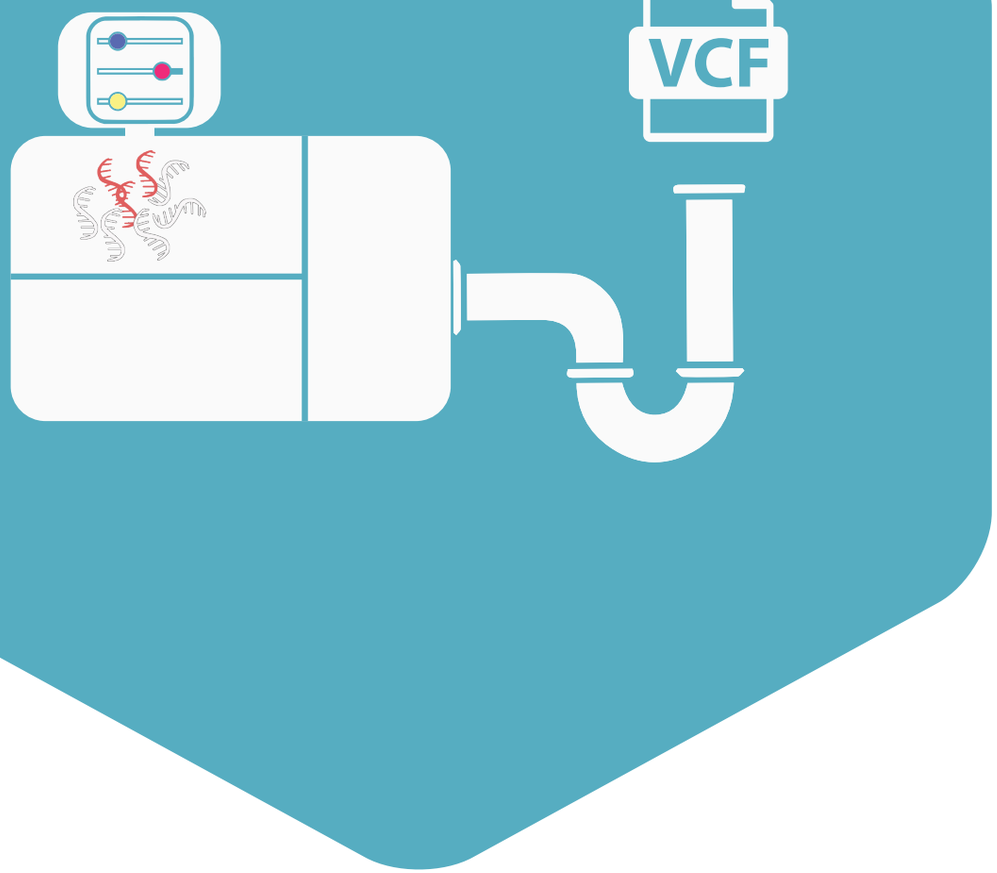


Starting point



SEQUENCING DESIGN & VARIANT CALLING PIPELINE

1



EUCEGENE PROJECT

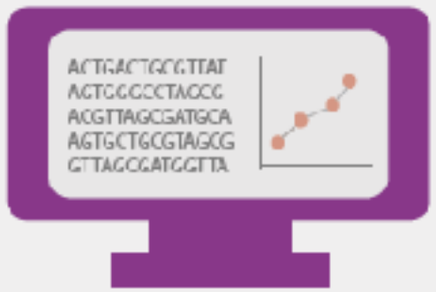
Improving AML Diagnosis and Therapies

46 CBF AML **public** RNA-Seq samples

Mutation type	No. Variants
Composite	9
Long Insertion	3
Short Deletion	2
Short Insertion	15
SNVs	58
Not reported	1

1 Get **high coverage** data from the same disorder with **published results**

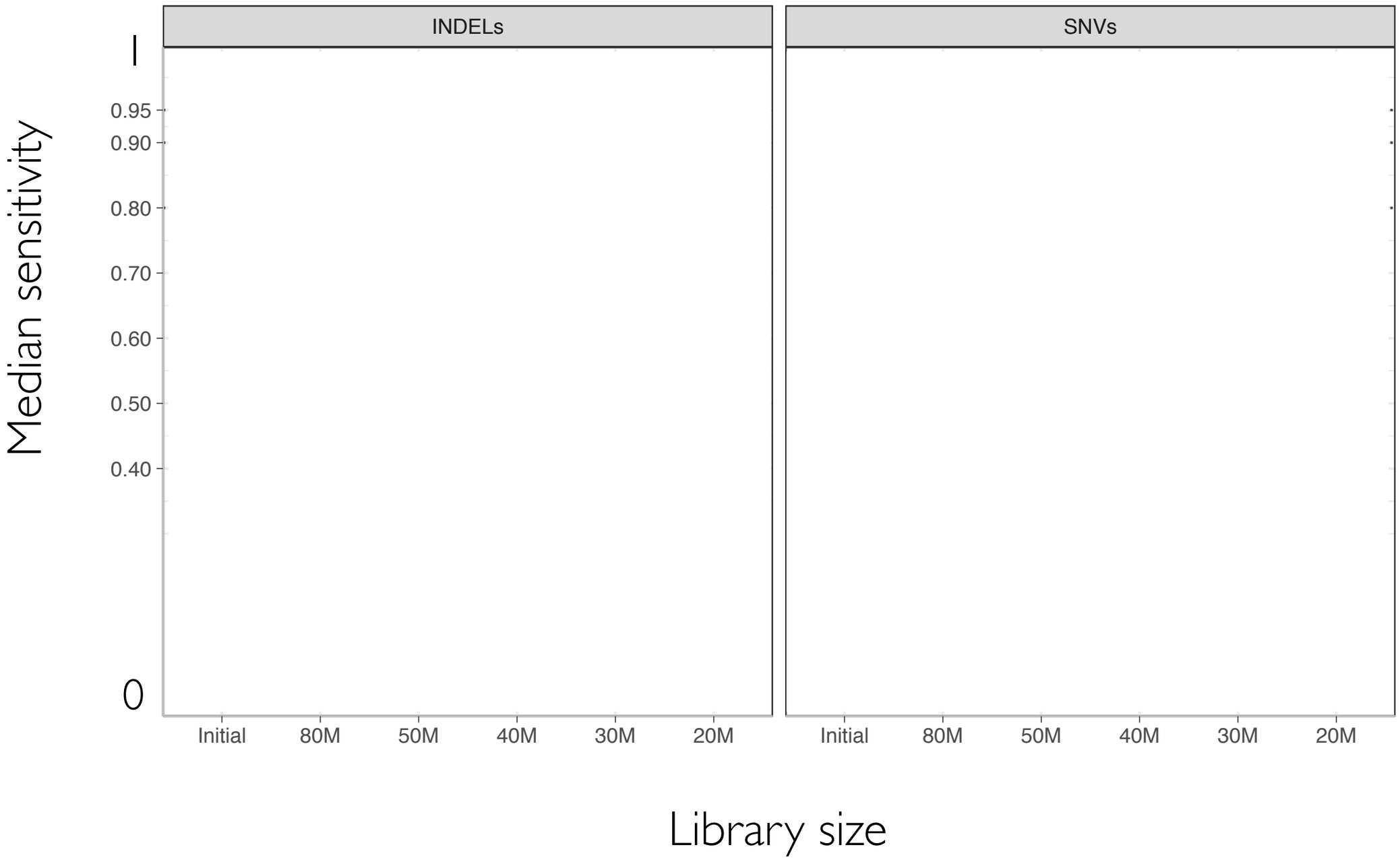
2 **Replicate** published results



3 **Compare** your results with the published ones

4 **Downsample**

Finding a suitable library size to call variants in RNA-Seq samples



caller
MuTect2
VarDict
VarScan2

Strategy
Strategy1
Strategy2

INDELs
SNVs

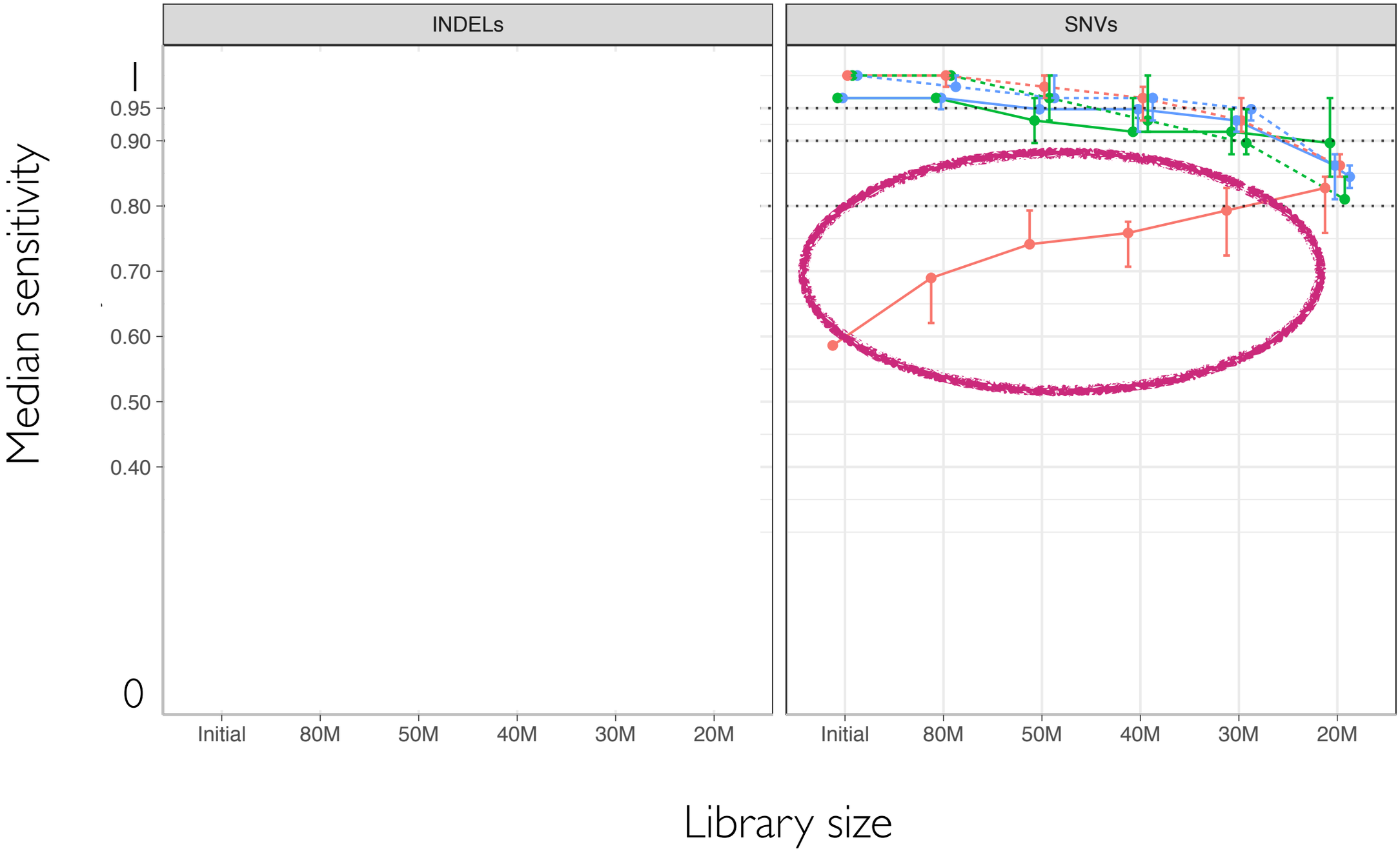
Median sensitivity

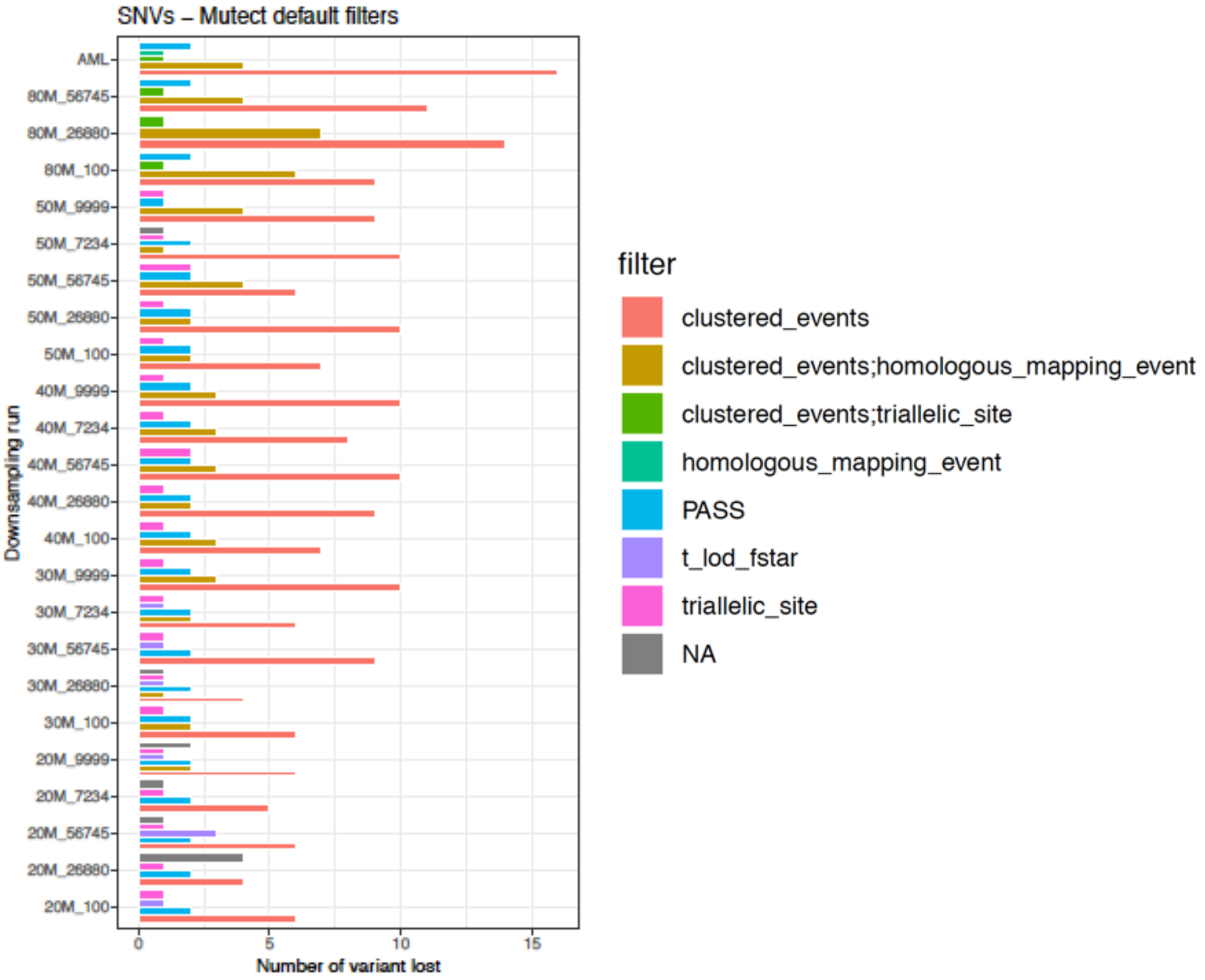
1
0.95
0.90
0.80
0.70
0.60
0.50
0.40
0

Initial
80M
50M
40M
30M
20M

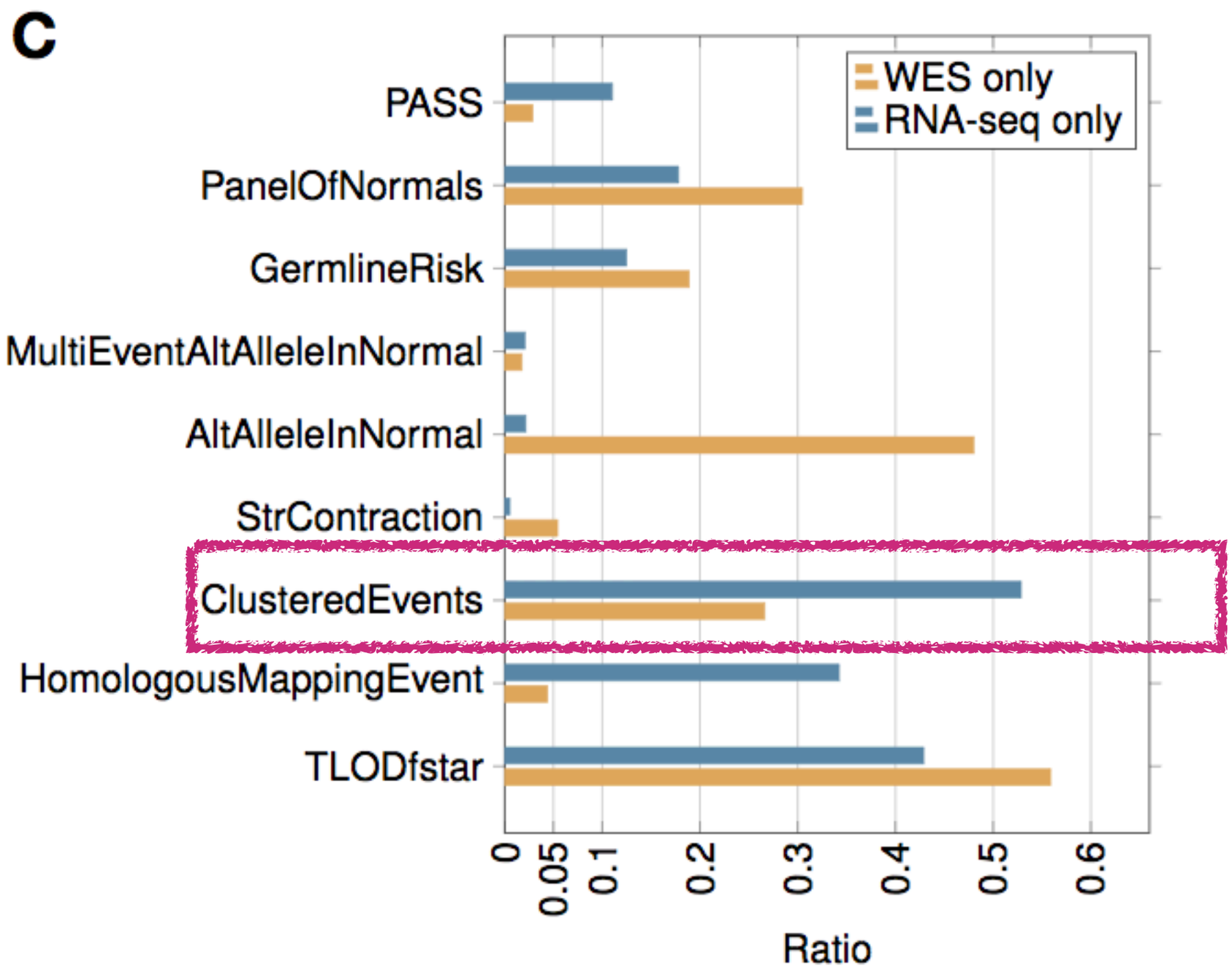
Library size

Getting better as library size decreases??




SNVs – Mutect default filters
AML
80M_56745
80M_26880
80M_100
50M_9999
50M_7234
50M_56745
50M_26880
50M_100
40M_9999
40M_7234
40M_56745
40M_26880
40M_100
30M_9999
30M_7234
30M_56745
30M_26880
30M_100
20M_9999
20M_7234
20M_56745
20M_26880
20M_100
Downsampling run
0
5
10
15
Number of variant lost
filter
clustered_events
clustered_events;homologous_mapping_event
clustered_events;triallelic_site
homologous_mapping_event
PASS
t_lod_fstar
triallelic_site
NA

Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data (Coudray et al 2018)



caller
MuTect2
VarDict
VarScan2

Strategy
Strategy1
Strategy2

Median with min/max sensitivity across different library sizes (min/max)

INDELs
SNVs

Median sensitivity

0.95
0.90
0.80
0.70
0.60
0.50
0.40
0

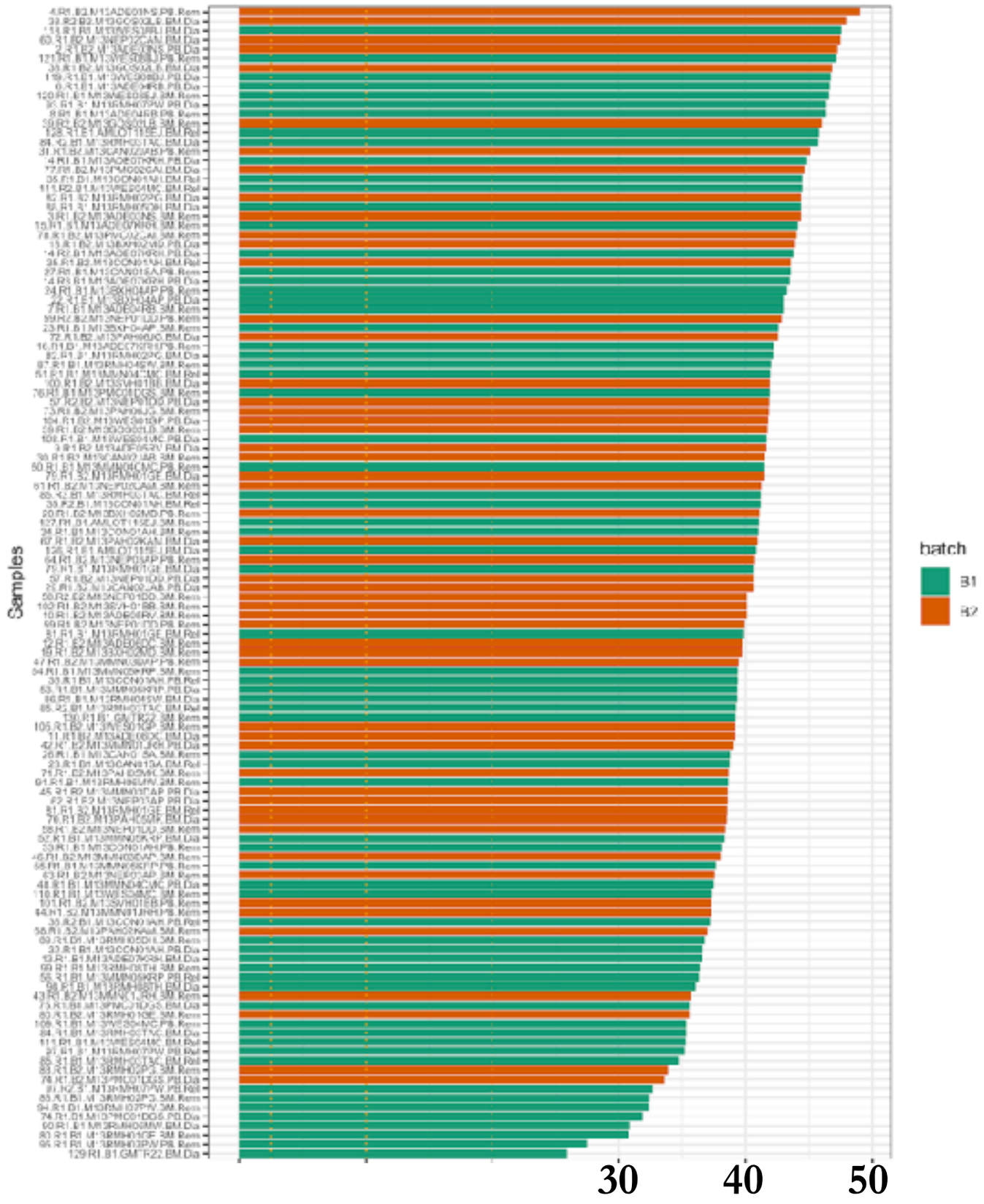
Initial
80M
50M
40M
30M
20M

Library size

Sequencing design and pipeline

- Library size: At least **30M PE** reads for each RNA samples

Uniquely mapped PE reads - ALLG



Decision about sequencing and variant calling

- Library size: At least 30M PE reads for each RNA samples

- Variant calling pipeline:
  - Call SNVs with VarScan, quicker than MuTect2

Decision about sequencing and variant calling

- Library size: At least 30M PE reads for each RNA samples
- Variant calling pipeline:
  - Call SNVs with VarScan, quicker than MuTect2
  - Call targeted INDELs with **VarDict** and **km**

Two pipelines

SNVs

VarScan²
(other callers are fine)

Get variants positions with a DNA caller → .vcf

Targeted INDEL calls

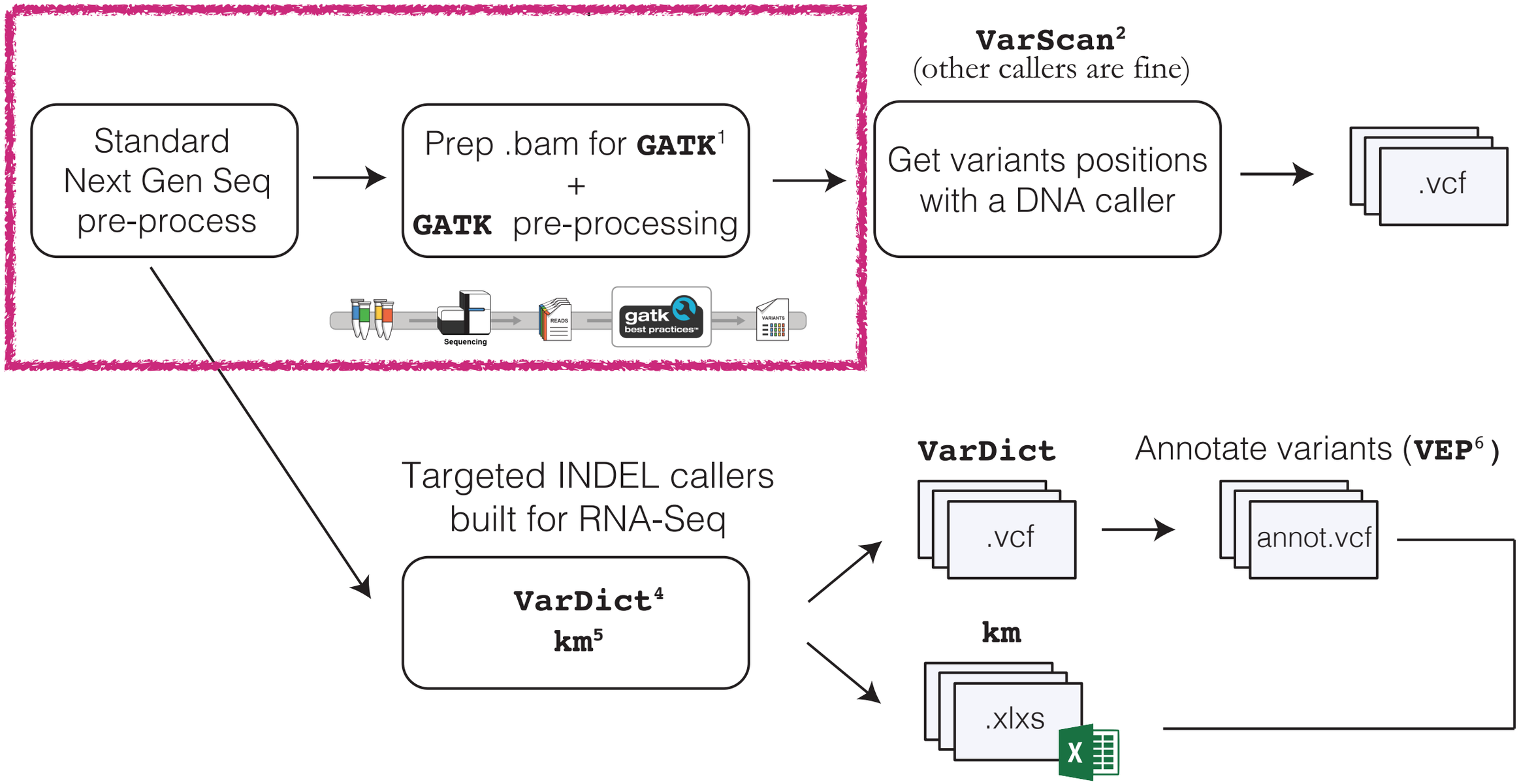
Targeted INDEL callers built for RNA-Seq

VarDict⁴
km⁵

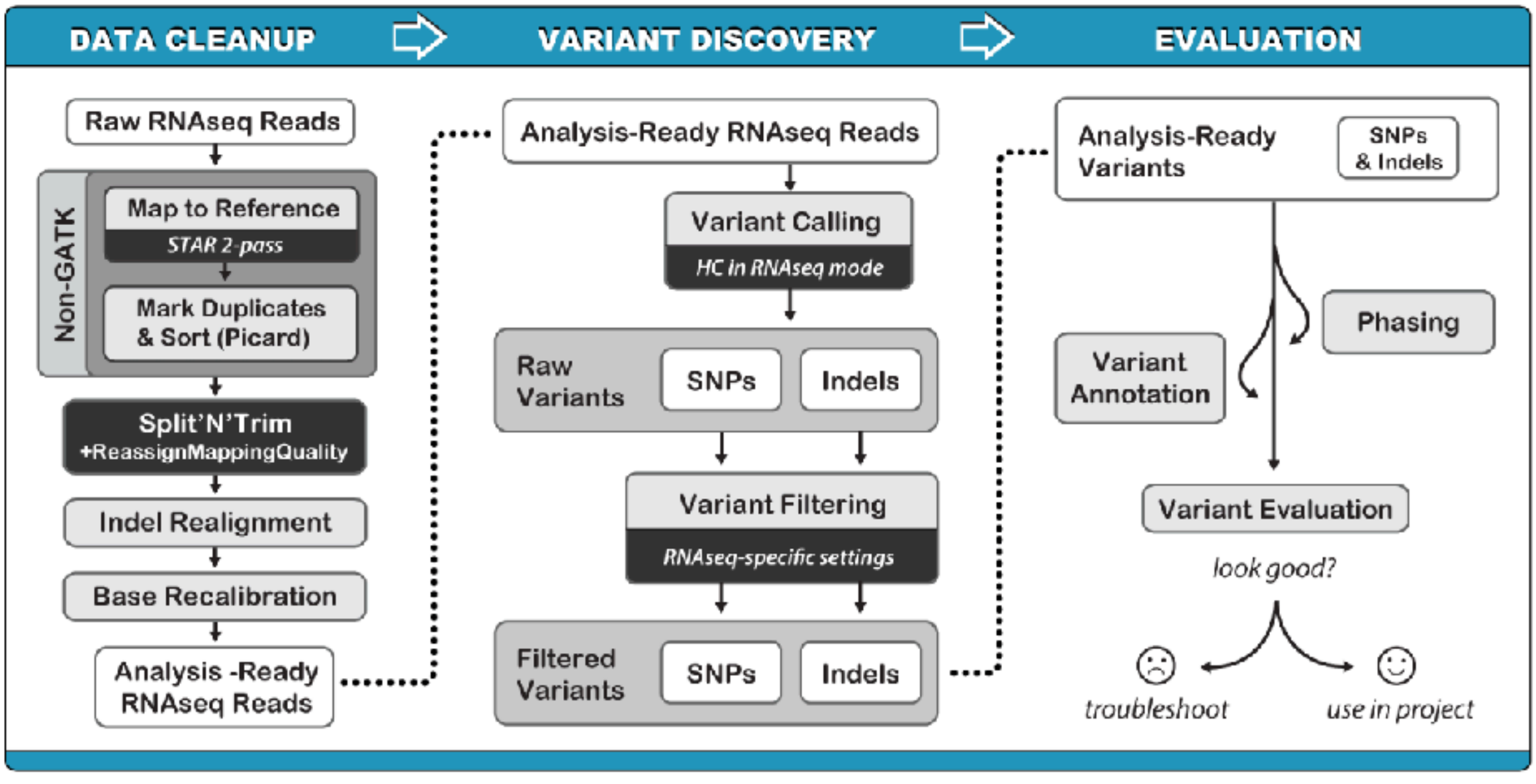
VarDict → .vcf → Annotate variants (**VEP⁶**) → annot.vcf

km → .xlxs

Pre-processing



GATK RNA-Seq best practices



https://software.broadinstitute.org/gatk/documentation/article.php?id=3891

Standard
Next Gen Seq
pre-process

Prep .bam for **GATK**[1]
+
GATK pre-processing

VarScan[2]
(other callers are fine)

Get variants positions
with a DNA caller

.vcf

Targeted INDEL callers
built for RNA-Seq

VarDict[4]
km[5]

VarDict

.vcf

Annotate variants (**VEP**[6])

annot.vcf

km

.xlxs

2

superFreq

mode = ‘RNA’

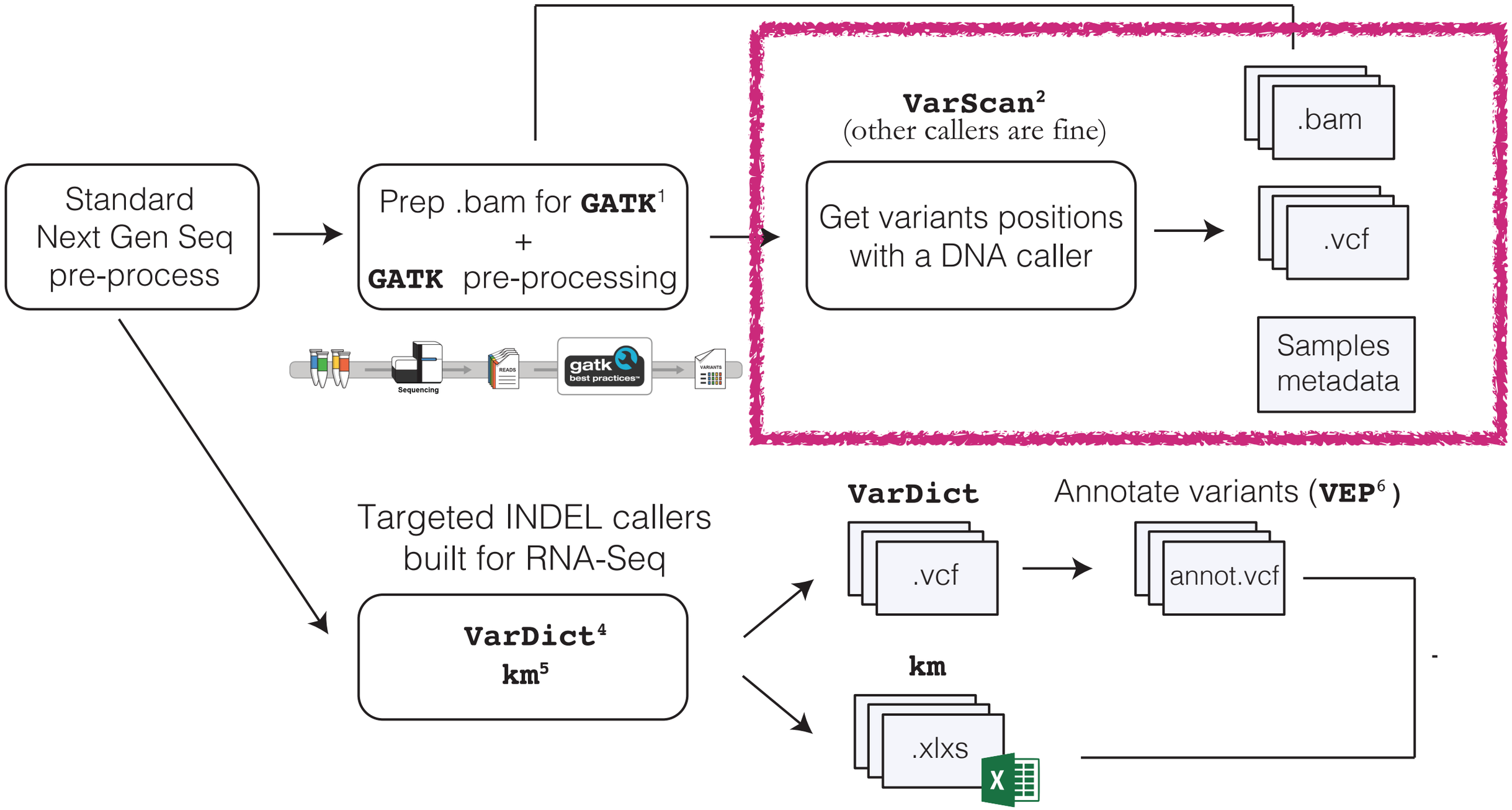
varikondo

COMBINE,
& DE-CLUTTER VARIANTS

SuperFreq: Integrated mutation detection and clonal tracking in cancer

Christoffer Flensburg, Tobias Sargeant, Alicia Oshlack, Ian Majewski

doi: https://doi.org/10.1101/380097



superFreq

https://github.com/ChristofferFlensburg/superFreq

Standard Next Gen Seq pre-process

Prep .bam for **GATK**[1] + **GATK** pre-processing

VarScan[2] (other callers are

Get variants pos with a DNA ca

.bam

.vcf

Samples metadata

superFreq[3]

`mode = 'RNA'`

Copy Number calls
SNVs
Clone tracking

Targeted INDEL callers built for RNA-Seq

VarDict[4]
km[5]

VarDict

.vcf

Annotate variants (**VEP**[6])

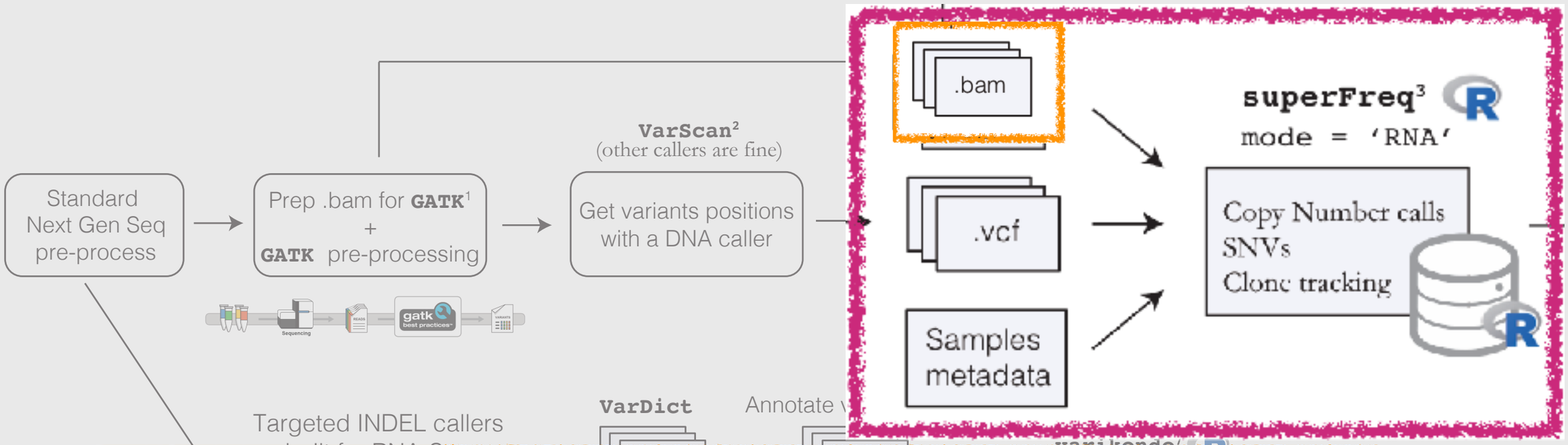
annot.vcf

km

.xlxs

superFreq

https://github.com/ChristofferFlensburg/superFreq



Includes a set of Panel Of Normal (PON) RNA-Seq

Standard
Next Gen Seq
pre-process

Prep .bam for **GATK**[1]
+
GATK pre-processing

VarScan[2]
(other callers are fine)

Get variants positions
with a DNA caller

.bam

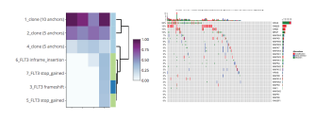
.vcf

Samples
metadata

superFreq[3]
`mode = 'RNA'`

Copy Number calls
SNVs
Clone tracking

Combine
standardise
visualise



Targeted INDEL callers
built for RNA-Seq

VarDict[4]
km[5]

VarDict

.vcf

Annotate variants (**VEP**[6])

annot.vcf

km

.xlxs

Combine
filter
curate

Simply…

`varikondo`

annaquaglieri16.github.io/varikondo/

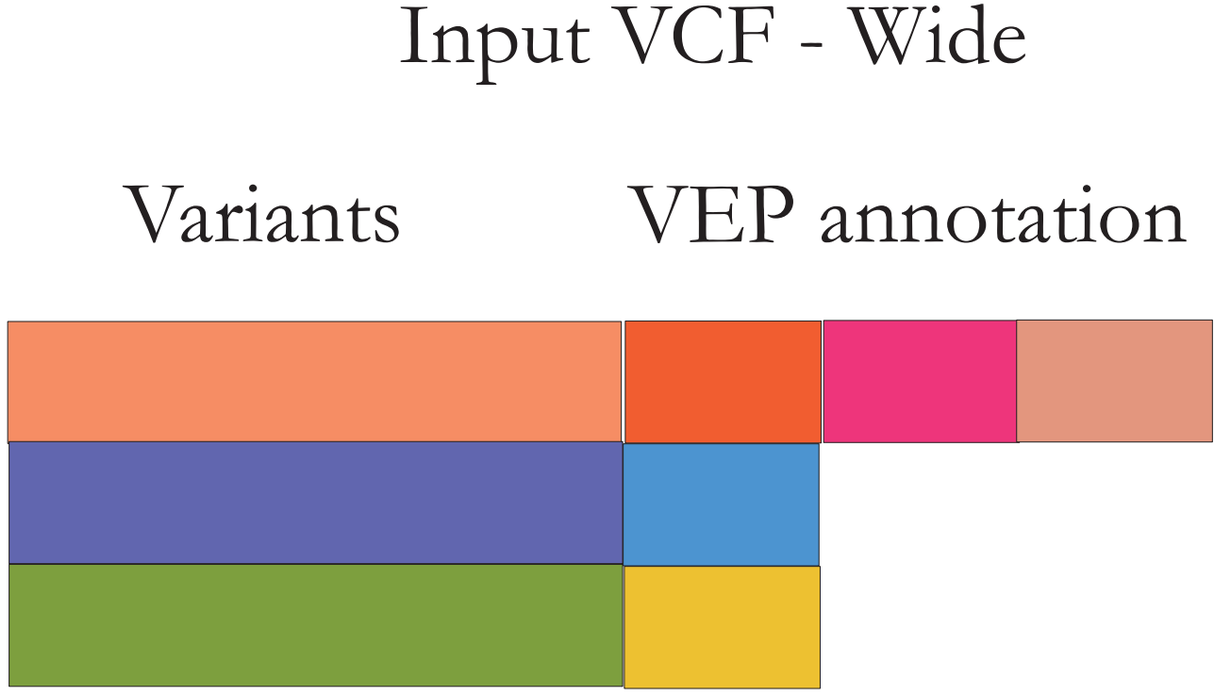
varikondo

`import_vcf(...,vep = TRUE)`

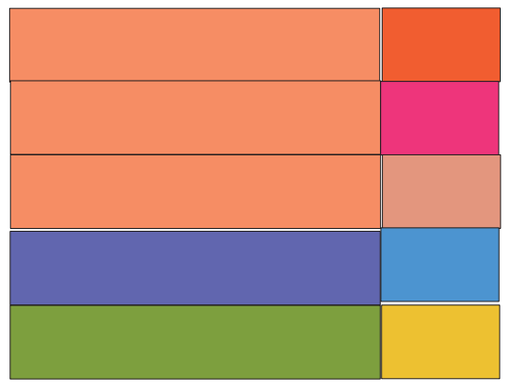
Input VCF - Wide

Variants

VEP annotation

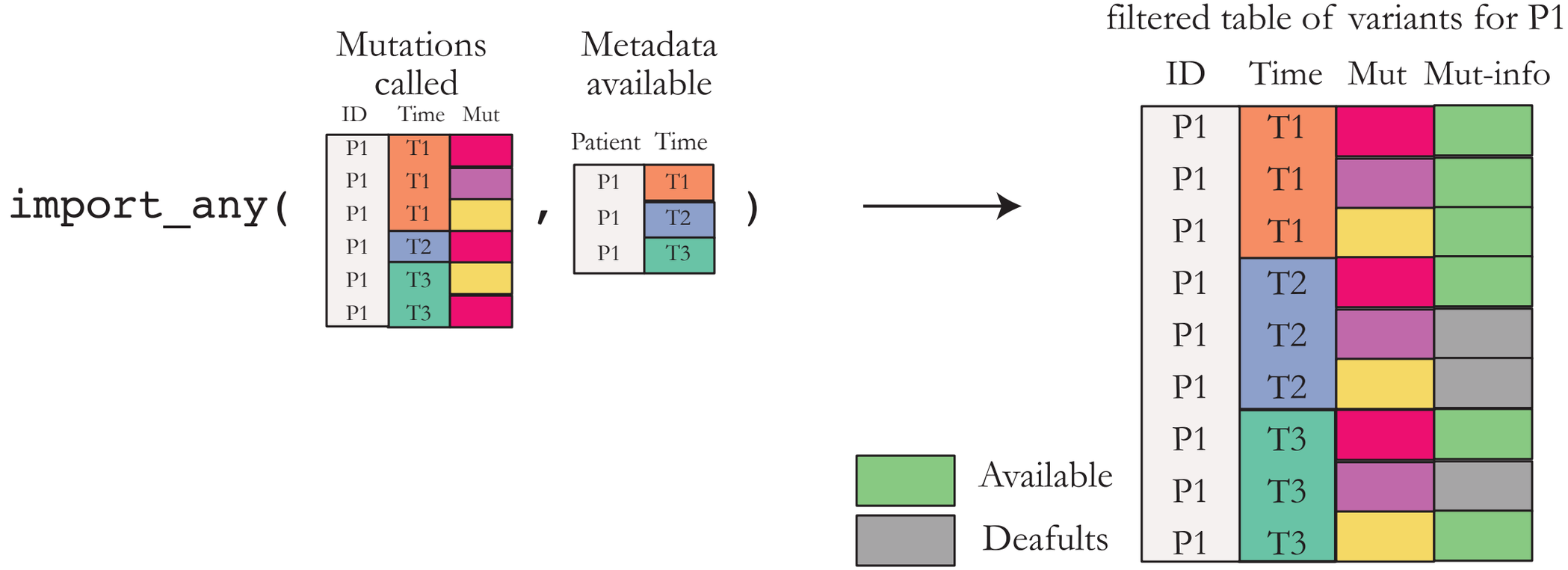


Output VCF - long



varikondo

import_any



Reference for import_any

More on INDEL filtering

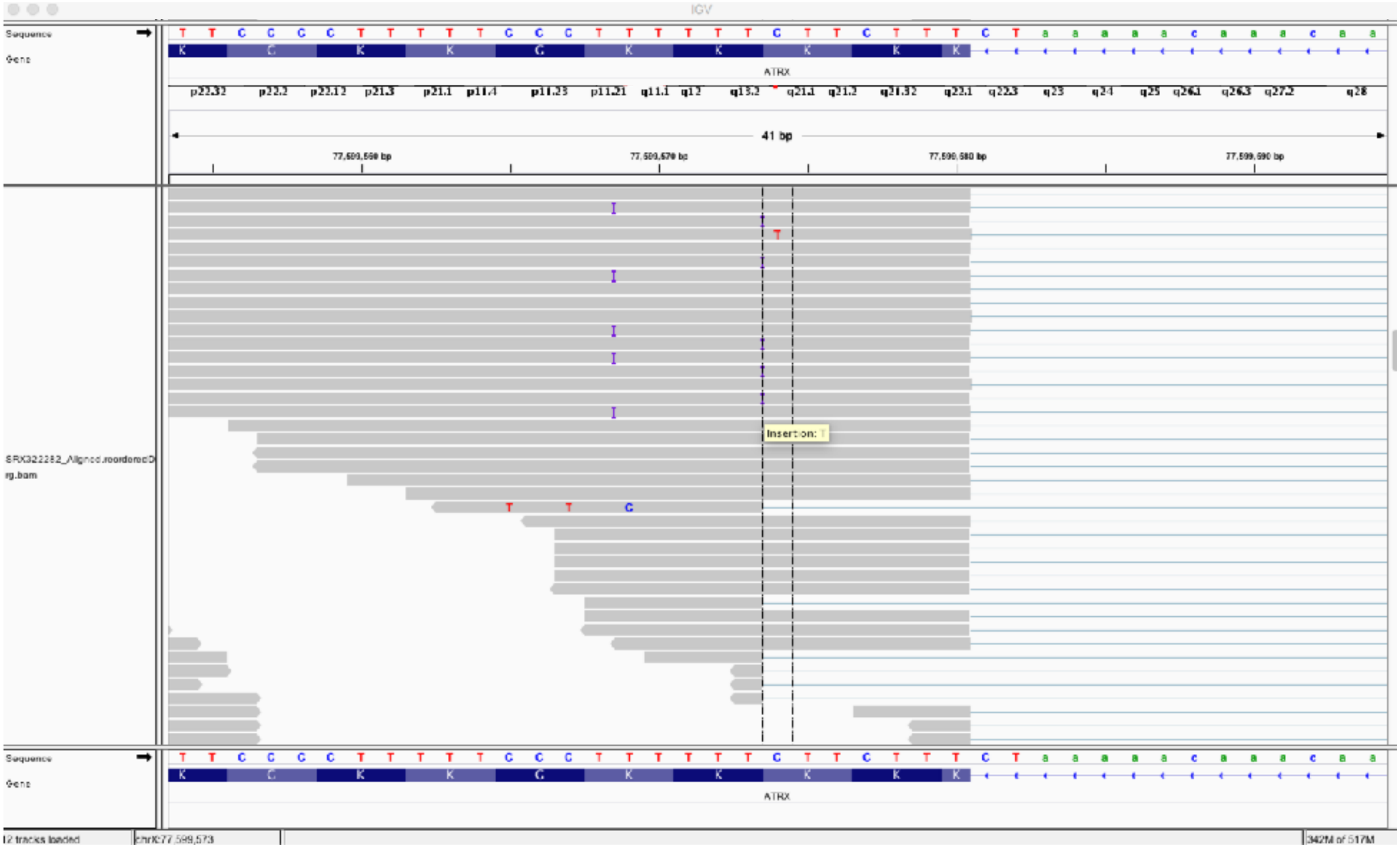
- Remove INDELs **called in normals**

More on INDEL filtering

- Remove INDELs **called in normals**

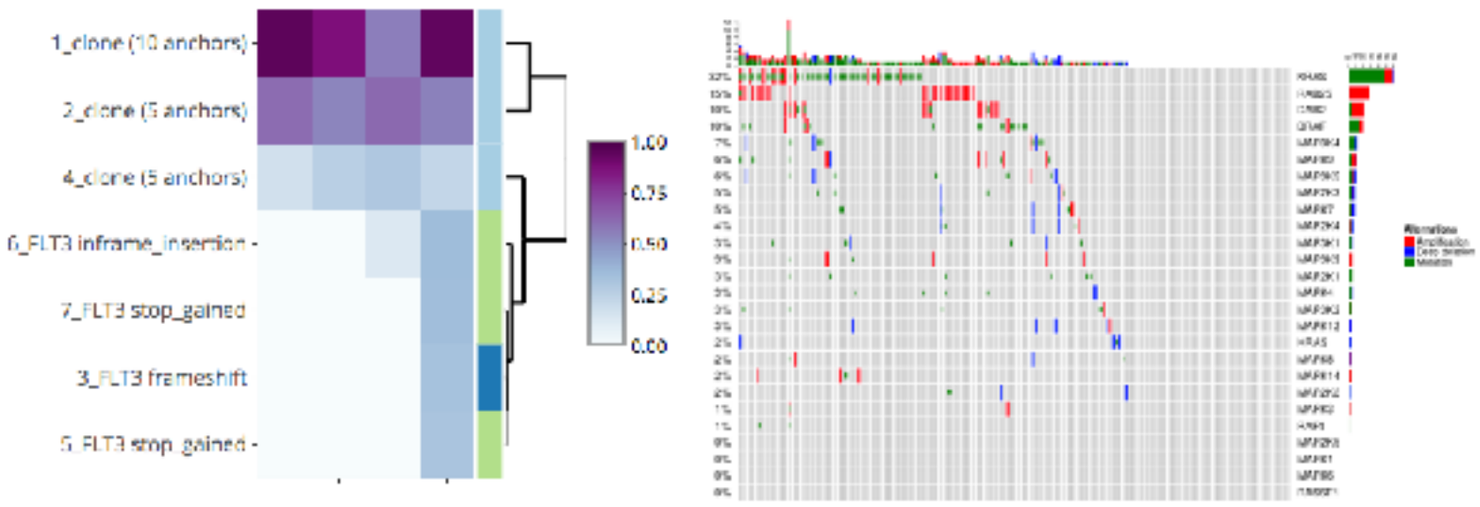
- Remove INDELs overlapping **repeat regions and homopolymers**

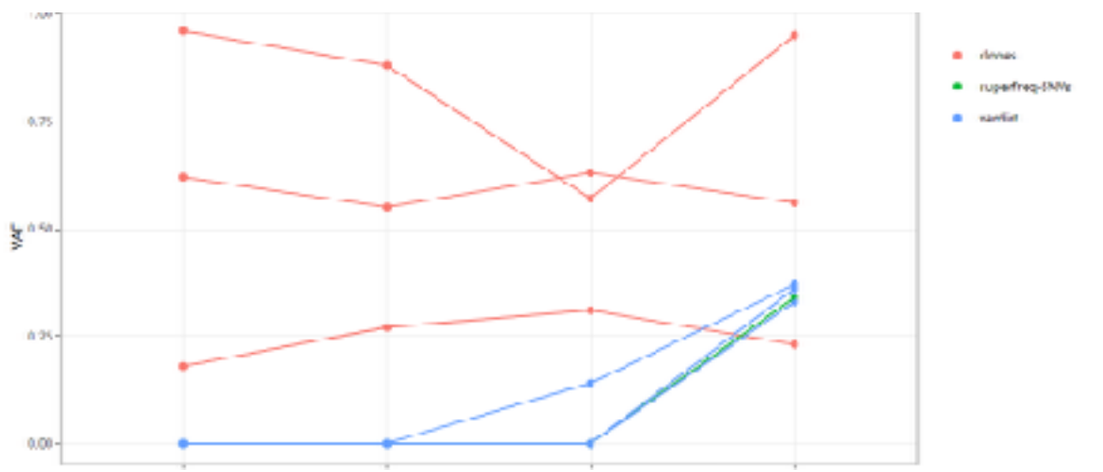
Homopolymers + In normals



Finally spark joy!

Explore variants





Limitations

Limitations

- Cannot detect mutations in non-expressed genes

Limitations

- Cannot detect mutations in non-expressed genes
- Non-sense mediated decay mutations, RNA gets "eaten" very quickly

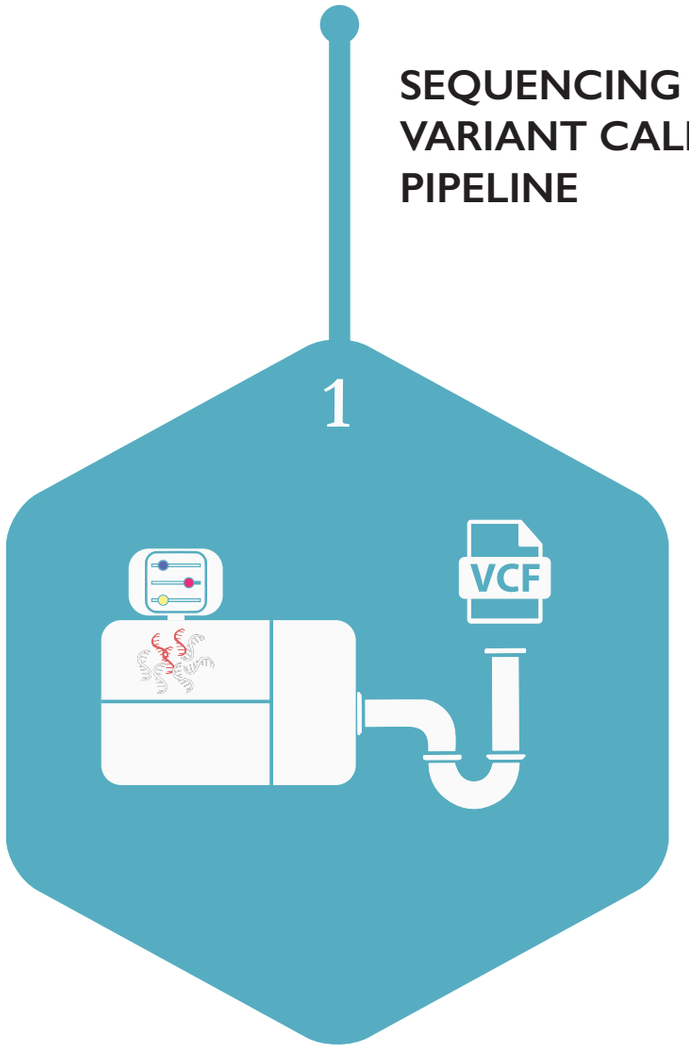
Limitations

- Expression varies on genes and we cannot detect mutations in non-expressed genes
- Non-sense mediated decay mutations, RNA gets "eaten" very quickly, decays and we can't detect the mutations!
- Allele specific expression. There is a DNA mutation but only one allele is expressed

Limitations

- Cannot detect mutations in non-expressed genes
- Non-sense mediated decay mutations, RNA gets "eaten" very quickly
- Allele specific expression.
- Less precise estimate of the variant allele frequency for one mutation

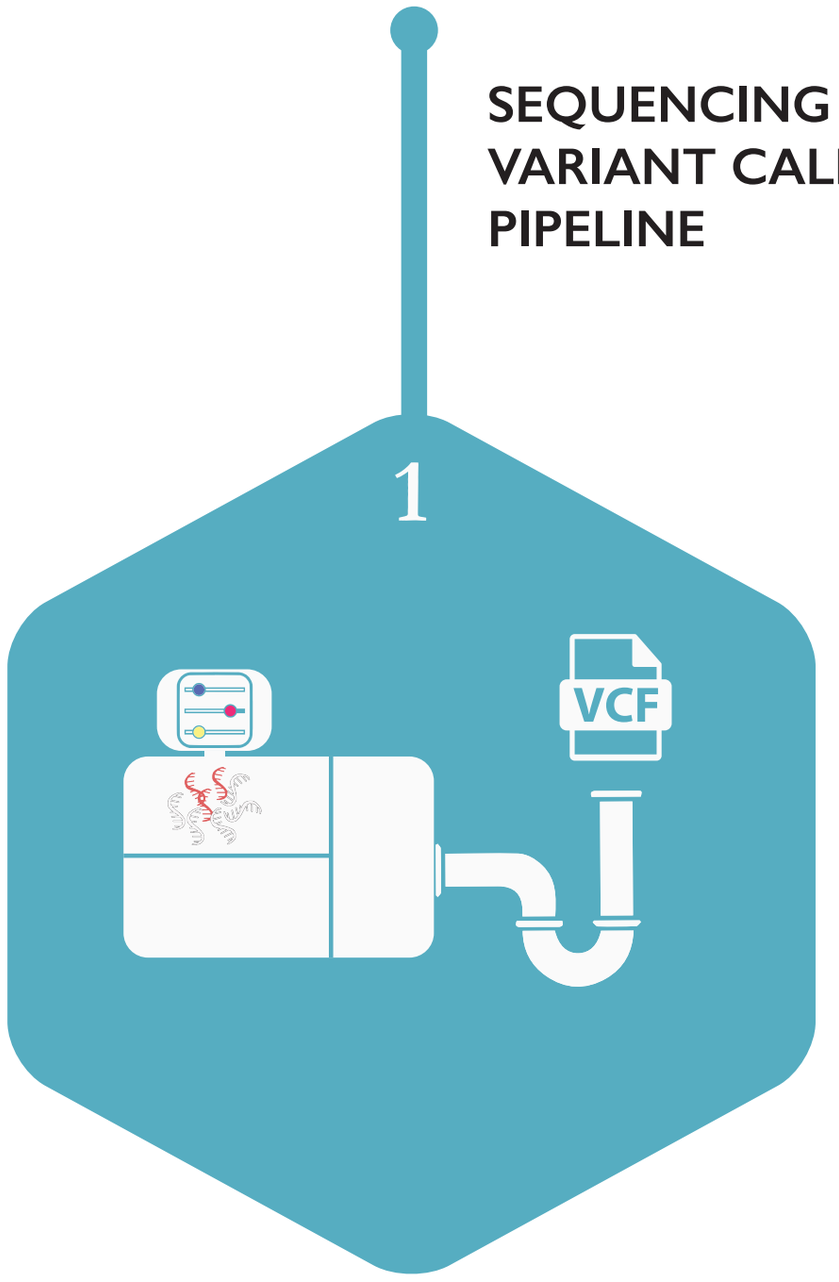
Summary

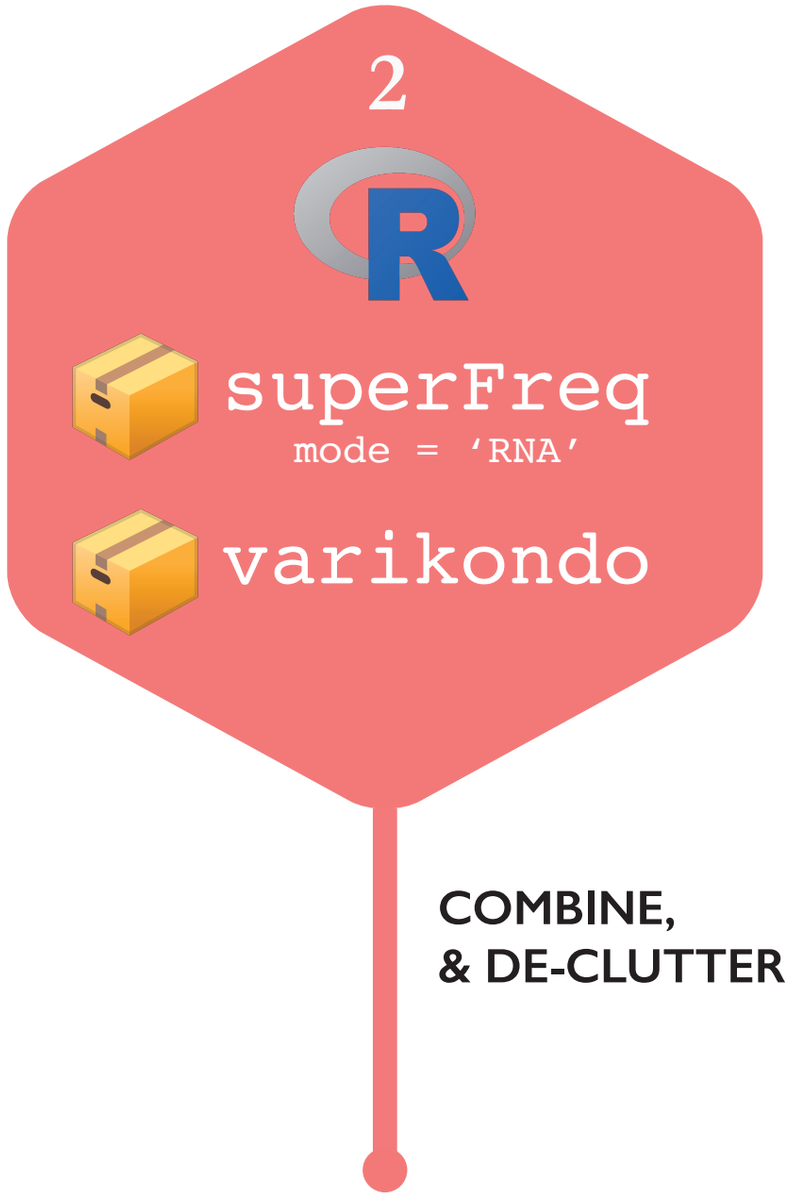


Variant calling pipeline functions
https://rna-mutation-calls.netlify.com/

Summary

SEQUENCING DESIGN &
VARIANT CALLING
PIPELINE





COMBINE,
& DE-CLUTTER VARIANTS

5

COMBINE HETEROGENOUS EXPRESSION DATA.

4

EXPLORE MUTATIONS

6

COMBINE MUTATION AND EXPRESSION RESULTS

Come to my completion seminar on the 19th August!

Acknowledgements

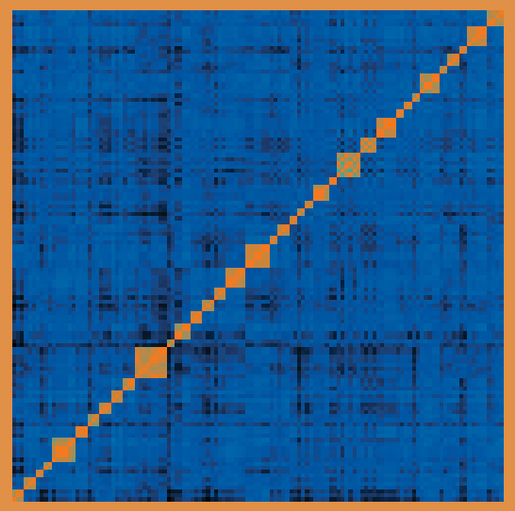
- Christoffer Flensburg
- Terry Speed
- Ian Majewki
- Stuart Lee
- Dharmesh Bhuva
- **Helpdesk**
- Earo Wang

Find material on `annaquaglieri16`

- https://github.com/annaquaglieri16/varikondo
- Variant calling pipeline functions https://rna-mutation-calls.netlify.com/

USE VARIANTS FOR QUALITY CONTROL

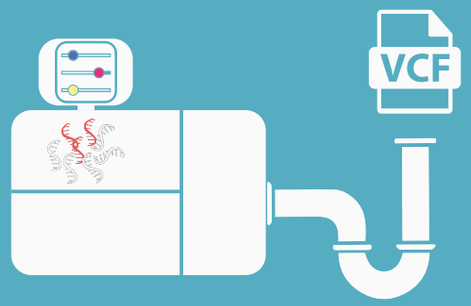
3



Summary

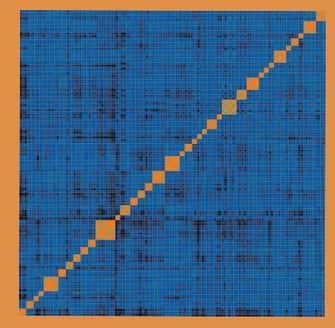
SEQUENCING DESIGN & VARIANT CALLING PIPELINE

1



2

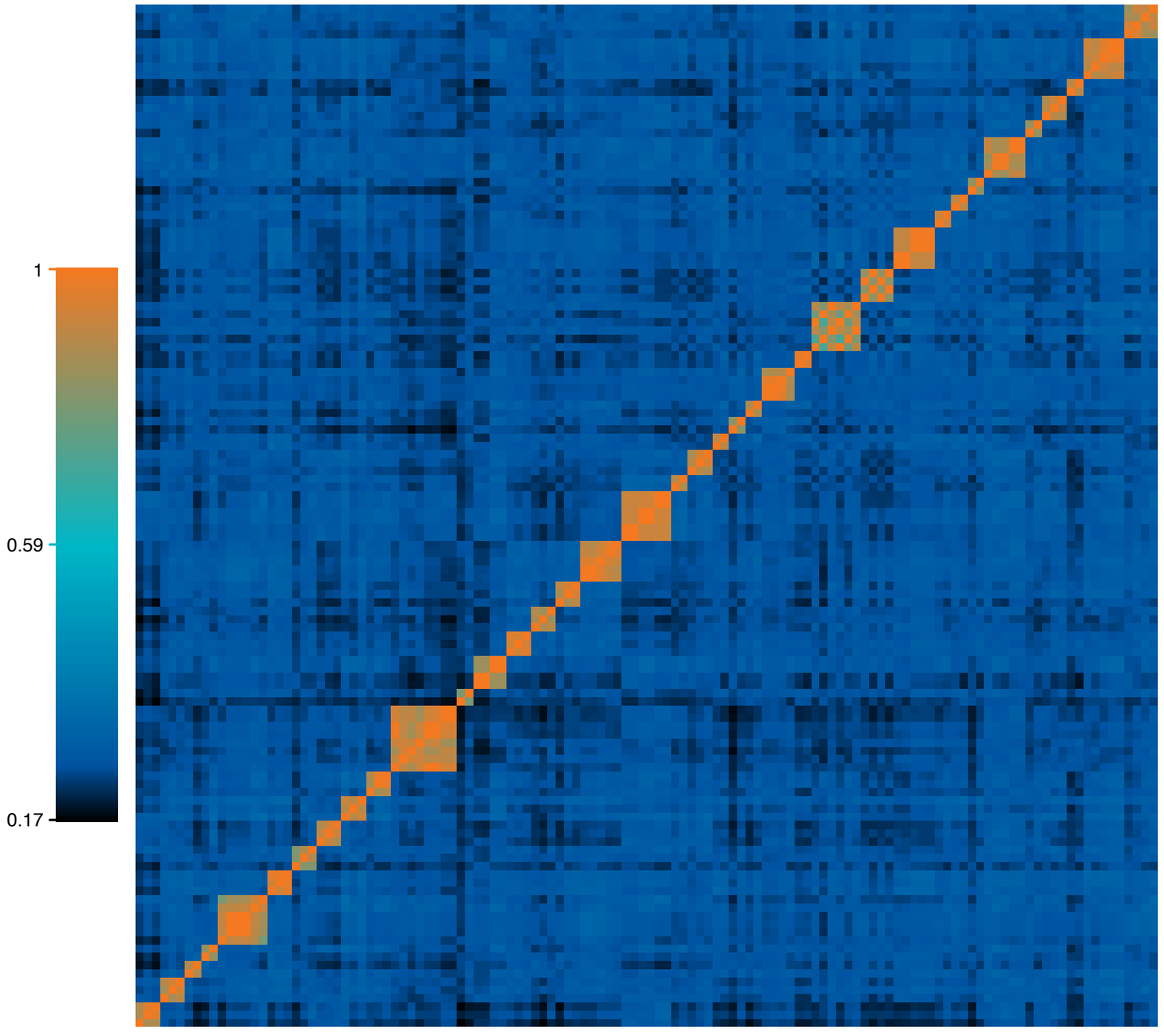
3



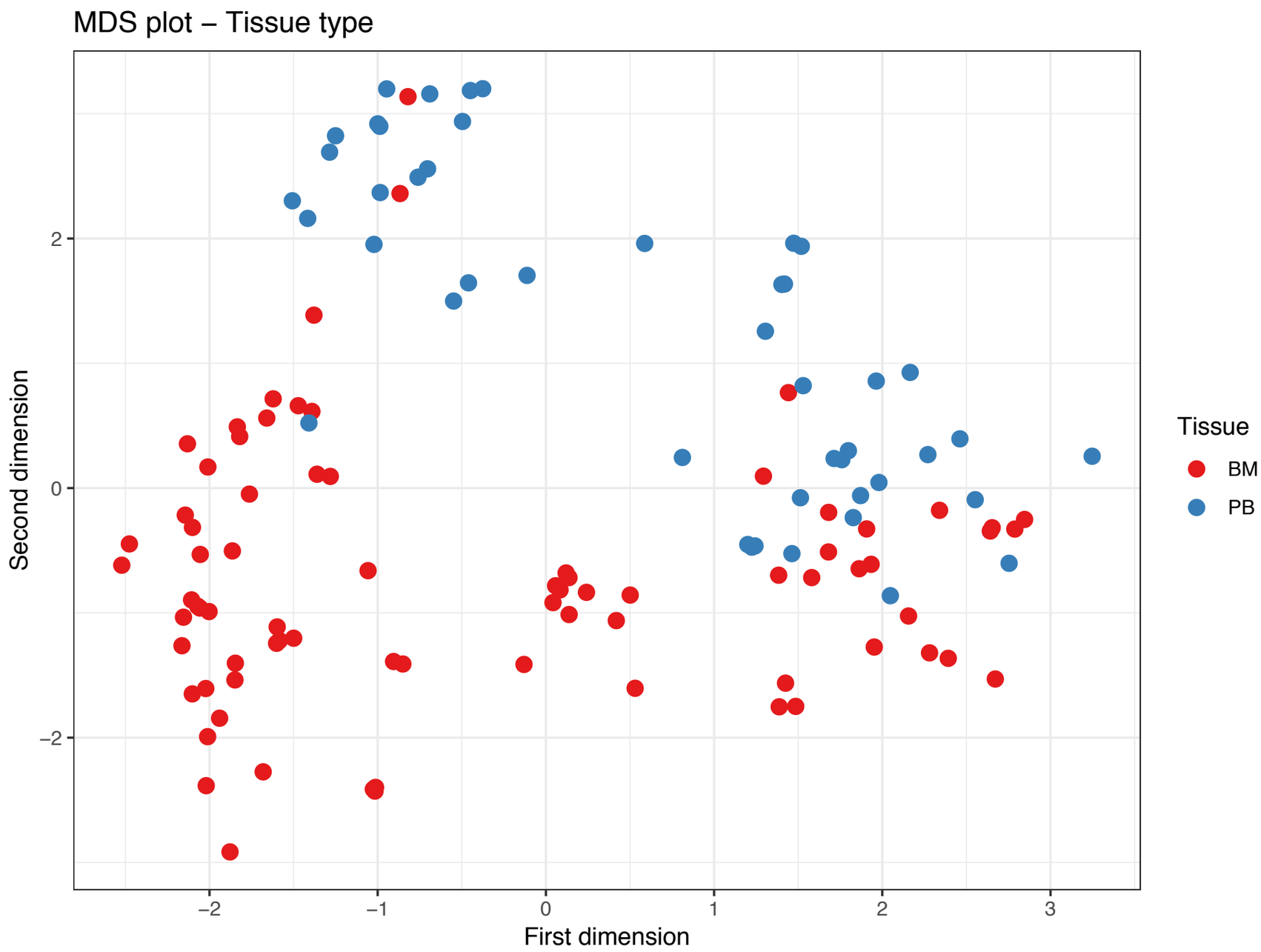
USE VARIANTS FOR QUALITY CONTROL

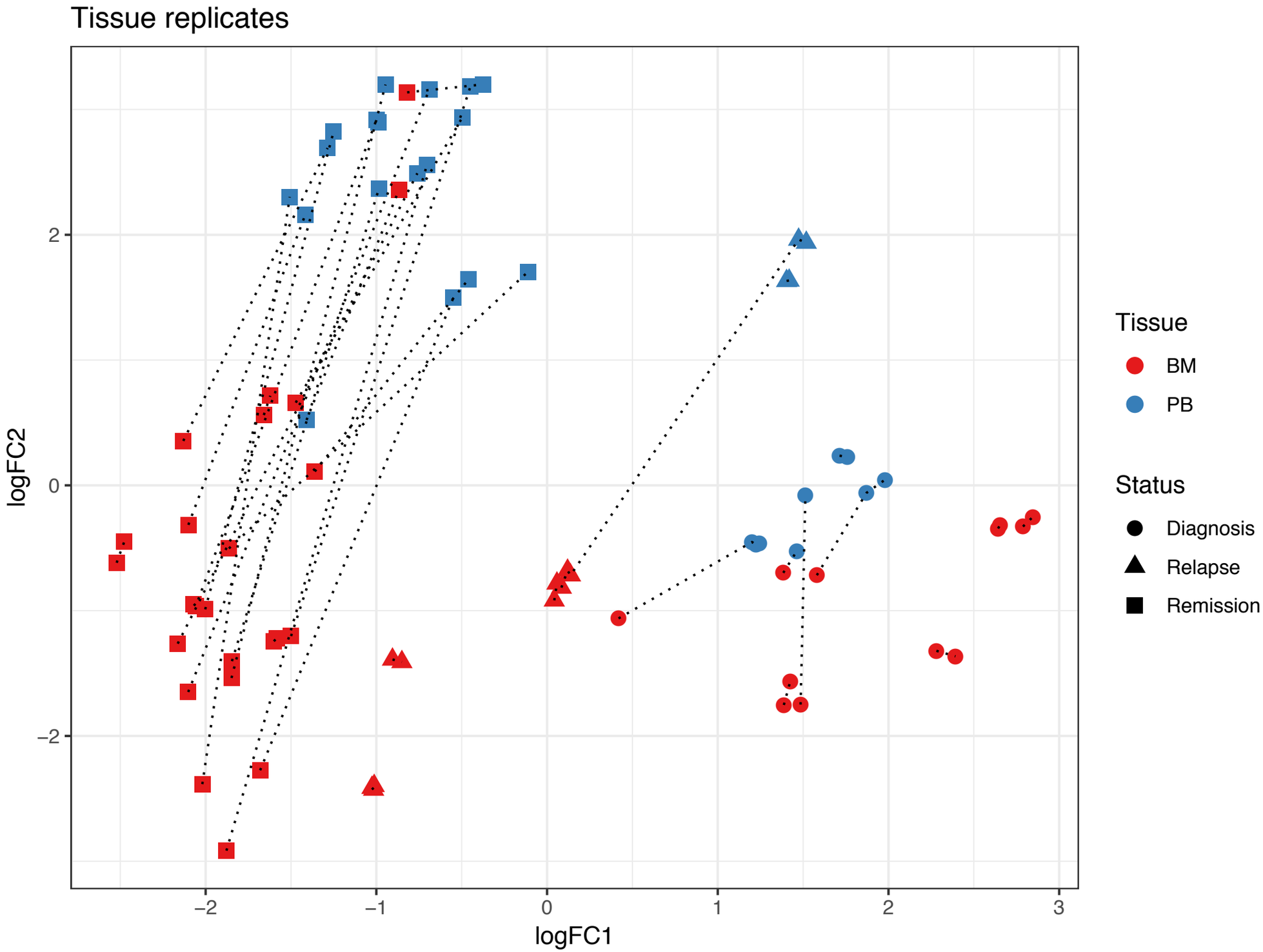
COMBINE, & DE-CLUTTER VARIANTS

Concordance of heterozygous SNPs

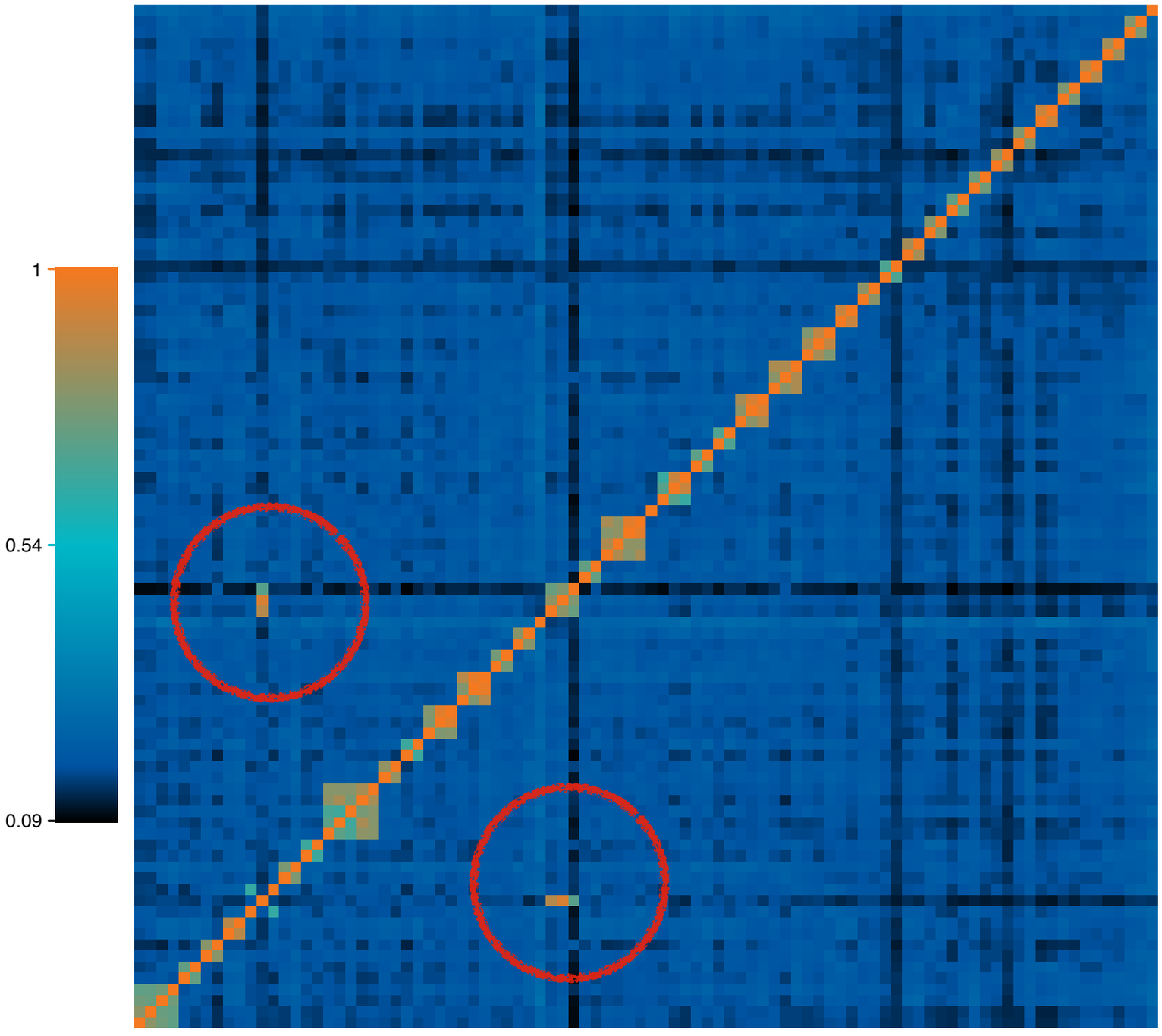


MDS plot using ALLG data

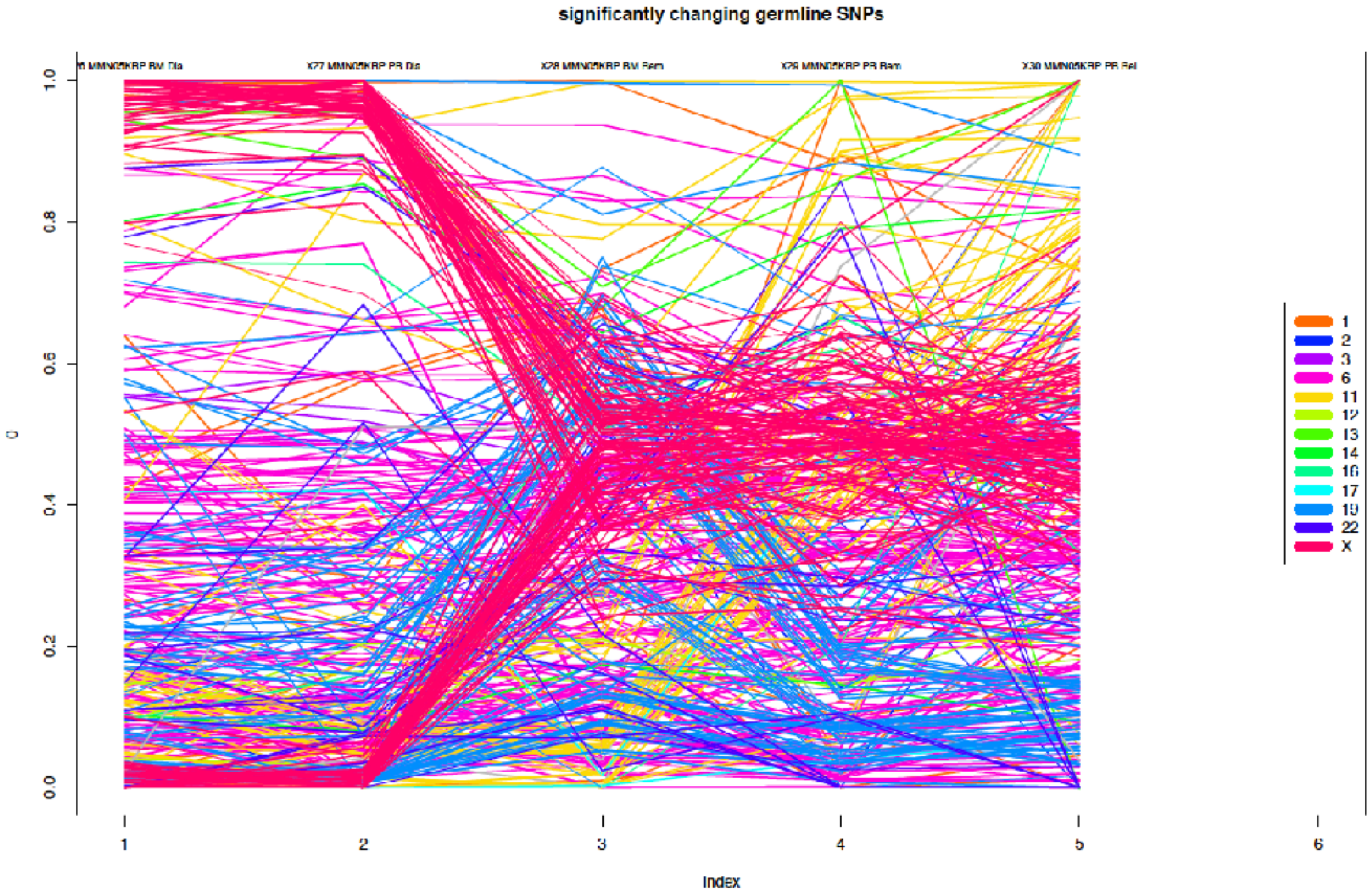



Tissue replicates
logFC2
2
0
−2
−2
−1
0
1
2
3
logFC1
Tissue
BM
PB
Status
Diagnosis
Relapse
Remission

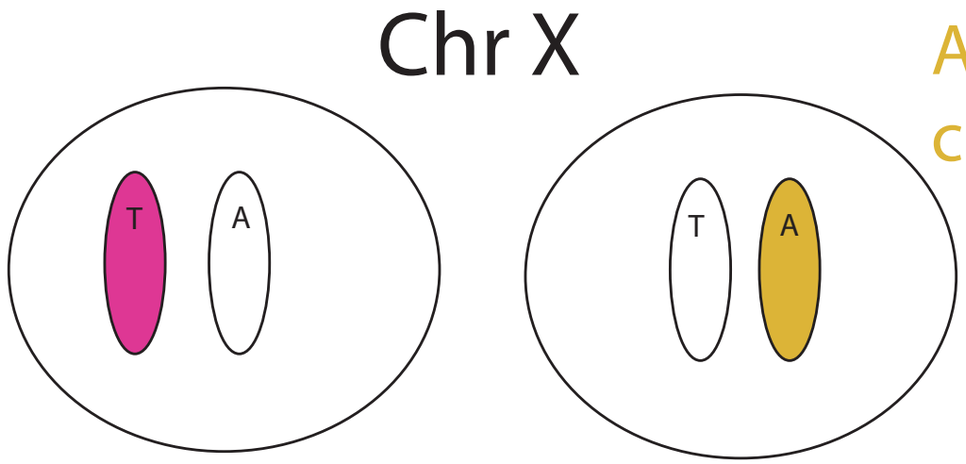
Sample swaps detected in a different RNA-Seq cohort



Curiosity!


significantly changing germline SNPs
X27 MMN05KRP PR Dis
X28 MMN05KRP RV Rem
X29 MMN05KRP PR Rem
X30 MMN05KRP PR Rel
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
Index
1
2
3
6
11
12
13
14
16
17
19
22
X

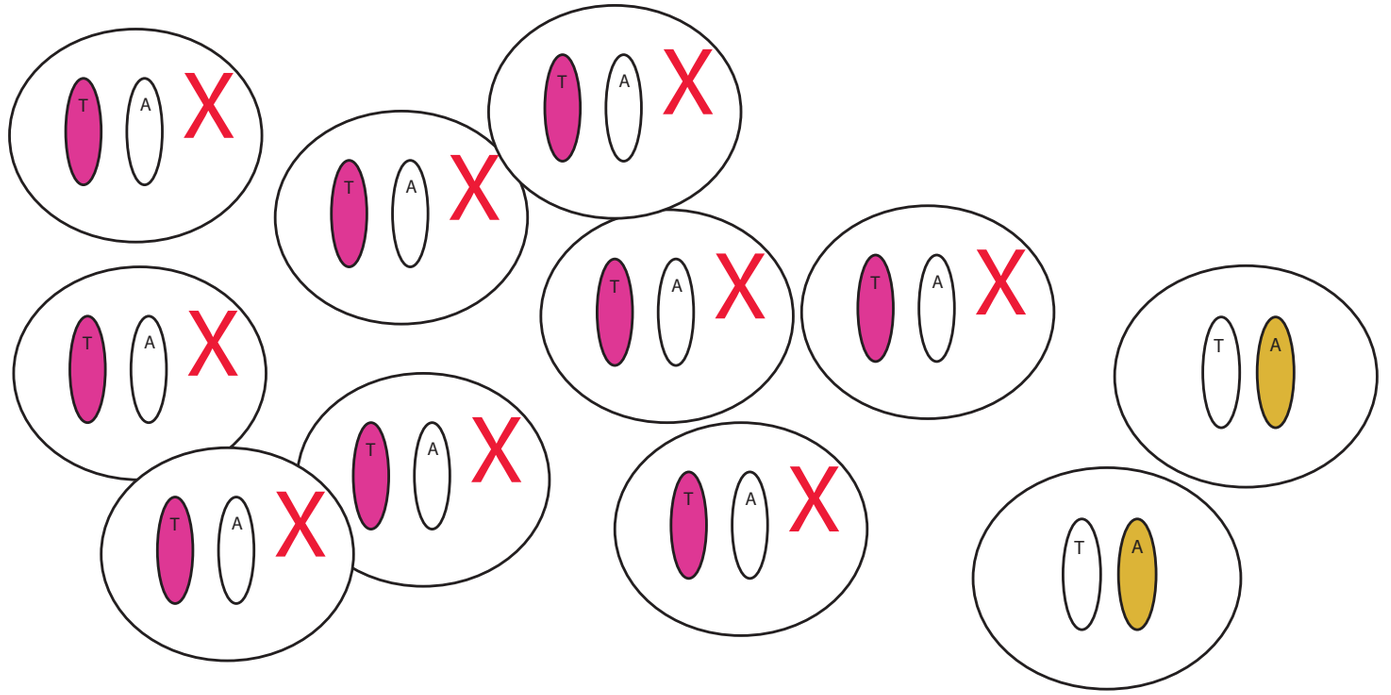
Activated chromosome



Activated chromosome

T A
T A
T A
T A
T A

Uncontrolled division
Cancer clone



T A
T A
T
T
T
T
T
T
T
T
T
T
T
T
T